



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<p>(51) International Patent Classification <sup>5</sup> : C12N 15/00, 15/67, 15/85 C12N 15/11, 5/10, C12Q 1/68</p>	<p>A1</p>	<p>(11) International Publication Number: <b>WO 92/13071</b> (43) International Publication Date: 6 August 1992 (06.08.92)</p>
<p>(21) International Application Number: PCT/US92/00692 (22) International Filing Date: 27 January 1992 (27.01.92) (30) Priority data: 646,664 28 January 1991 (28.01.91) US (71) Applicant: MASSACHUSETTS INSTITUTE OF TECHNOLOGY [US/US]; 77 Massachusetts Avenue, Cambridge, MA 02139 (US). (72) Inventors: BUCKLER, Alan, J. ; 133 Park Street, Brookline, MA 02146 (US). CHANG, David, D. ; 20 Webster Street, Apt. 403, Brookline, MA 02146 (US). SHARP, Phillip, A. ; 119 Grasmere Street, Newton, MA 02158 (US). HOUSMAN, David, E. ; 64 Homer Street, Newton, MA 02159 (US).</p>		<p>(74) Agents: GRANAHAH, Patricia et al.; Hamilton, Brook, Smith &amp; Reynolds, Two Militia Drive, Lexington, MA 02173 (US). (81) Designated States: AT (European patent), BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent).  Published <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i></p>
<p>(54) Title: METHOD OF EXON AMPLIFICATION</p> <p>(57) Abstract</p> <p>A method of exon amplification, which is useful for fast and efficient isolation of a coding sequence from complex mammalian genomic DNA. Fragments of genomic mammalian DNA are inserted into an intron contained within a splicing plasmid, resulting in a splicing plasmid construct. The construct is introduced into an appropriate host cell, resulting in replication of and transcription from the construct. The transcripts are processed into mature RNA. If an exon is present in the genomic DNA fragment contained in the plasmid intron, the splice sites of the DNA insert can be paired with 5' and 3' splice sites provided in the splicing construct. Mature RNA, which contains transcribed exons from the genomic DNA, is isolated, amplified via RNA-based PCR, and subsequently cloned.</p> <pre> graph TD     A[Splicing plasmid construct: SV40, 5' ss, cloning site, 3' ss, pA] --&gt; B[genomic fragment containing exon]     A --&gt; C[transfection by electroporation into COS cells]     B --&gt; C     C --&gt; D[cytoplasmic RNA isolation]     D --&gt; E[cDNA synthesis RT: AAAAAA]     E --&gt; F[PCR amplification]     F --&gt; G[amplified product containing introduced exon]   </pre>		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FI	Finland	ML	Mali
AU	Australia	FR	France	MN	Mongolia
BB	Barbados	GA	Gabon	MR	Mauritania
BE	Belgium	GB	United Kingdom	MW	Malawi
BF	Burkina Faso	GN	Guinea	NL	Netherlands
BG	Bulgaria	GR	Greece	NO	Norway
BJ	Benin	HU	Hungary	PL	Poland
BR	Brazil	IE	Ireland	RO	Romania
CA	Canada	IT	Italy	RU	Russian Federation
CF	Central African Republic	JP	Japan	SD	Sudan
CG	Congo	KP	Democratic People's Republic of Korea	SE	Sweden
CH	Switzerland	KR	Republic of Korea	SN	Senegal
CI	Côte d'Ivoire	LI	Liechtenstein	SU	Soviet Union
CM	Cameroon	LK	Sri Lanka	TD	Chad
CS	Czechoslovakia	LU	Luxembourg	TG	Togo
DE	Germany	MC	Monaco	US	United States of America
DK	Denmark	MG	Madagascar		
ES	Spain				

-1-

METHOD OF EXON AMPLIFICATIONDescriptionBackground

Understanding the molecular basis of human  
05 genetic disorders and corresponding genotypes in  
other mammalian genomes requires methods for the  
identification of coding sequences in target  
chromosomal regions. Current methods which are used  
for this are both inefficient and tedious. The  
10 strategy used most frequently involves the screening  
of short genomic DNA segments for sequences which  
are evolutionarily conserved (Monaco, A.P. et al.,  
Nature (London) 323:646-650 (1986); Page, D.C. et  
al., Cell 51:1091-1104 (1987); Rommens, J.M. et al.,  
15 Science 245:1059-1065 (1989); Call, K.M. et al.,  
Cell 60:509-520 (1990). Although successful, there  
are several limitations to this approach. Sequences  
contained in mRNA do not necessarily show such  
cross-species conservation. To avoid interference  
20 with repetitive sequences, short segments of cloned  
genomic DNA must be tested individually as probes in  
Southern blotting experiments, a time consuming  
procedure. Finally, this standard approach does not  
directly lead to identification of transcribed  
25 segments of the cloned genomic DNA, but requires  
isolation of cDNAs corresponding to the transcribed  
region. This assumes knowledge of specific tissue  
expression of a given transcript, as well as  
adequate expression levels in the RNA used for  
30 construction of the cDNA library. Alternative

-2-

strategies for gene isolation involve sequencing and analysis of large segments genomic DNA for the presence of open reading frames (Fearon, E.R. et al., Science 247:49-56 (1990), and cloning of  
05 hypomethylated CpG islands, which are signposts of 5'-ends of transcription units (Bird, A., Nature (London) 321:209-213 (1986). These methods also do not provide a direct means of purifying coding sequences from genomic DNA. At present, an  
10 efficient, sensitive method of isolating coding sequences from complex genomic DNA is not available.

#### Summary of the Invention

The present invention relates to a method of exon amplification, which is useful for fast and  
15 efficient isolation of a coding sequence from complex mammalian genomic DNA. The present method is based upon the ability of RNA transcripts to undergo processing events which require the presence of splice recognition and branchpoint sequences. In  
20 the present method of isolating a coding sequence or exon from mammalian genomic DNA, fragments of genomic mammalian DNA are inserted into an intron contained within an in vivo splicing plasmid, resulting in production of an in vivo splicing  
25 plasmid construct, in which mammalian genomic DNA is present within a plasmid intron. The constructs are introduced (e.g., transfected by electroporation) into an appropriate host cell type, such as COS7 cells, in which they replicate. Host cells  
30 containing the in vivo splicing plasmid constructs are maintained under conditions appropriate for

-3-

replication of and transcription from the in vivo splicing plasmid constructs, resulting in production and in vivo processing of RNA in the host cells. If an exon is present in the genomic DNA fragment

05 contained within the plasmid intron, the vector splice sites are paired with splice sites of the DNA insert. The resulting mature RNA contains the previously unidentified exons (is a mature RNA transcript of the genomic DNA). The mature RNA

10 containing the exons is isolated and can then be amplified via RNA-based PCR, and subsequently cloned. Alternatively, the PCR product can be purified and sequenced directly. For example, mature RNA obtained from host cells in which in vivo

15 plasmid constructs have replicated can be screened with an appropriate probe, such as an anti-sense oligonucleotide to identify mature RNA transcripts of the exon. RNA transcripts of the exon are amplified using an RNA-based amplification method

20 (e.g., RNA-based PCR), thereby producing cDNA of the mammalian genomic DNA. cDNA produced in this manner is purified (to separate it from other cDNA present) and the purified cDNA is directly sequenced or digested with at least one restriction enzyme which

25 recognizes a restriction site or sites present in the splicing vector or in the PCR product. The digested purified cDNA is introduced into the cloning vector (e.g., pBluescript/11SK+, Stratagene) and cloned. The cloned DNA can be sequenced using

30 known methods.

In one embodiment of the present invention, genomic fragment(s) having compatible ends is cloned

-4-

into the in vivo splicing plasmid pSPL1 at a BamHI site created within the HIV-tat intron (See Figure 1a). The clone is then transiently transfected by electroporation into COS7 cells, after which

05 amplification of the plasmid occurs by virtue of the SV40 origin of replication. High levels of transcription are facilitated by the SV40 promoter, resulting in production of RNA containing the introduced genomic sequence. If the genomic

10 sequence contains an exon in the proper orientation, processing occurs in such a manner that the exon is retained in the mature RNA, flanked by HIV tat and  $\beta$ -globin exon sequences. Subsequently, cytoplasmic RNA is isolated from the COS7 cells, and subjected

15 to RNA-based PCR analysis using oligodeoxynucleotides which hybridize to the flanking  $\beta$ -globin sequences. The amplified product contains the introduced exon sequence, and can be analysed by cloning and sequencing or by direct PCR sequencing.

20 The present method has been used to isolate exon sequences from cloned genomic fragments known to contain exon sequences (i.e., fragments of a mouse clone known to contain exon sequences of the murine Na,K-ATPase  $\alpha$ 1 subunit gene). The present

25 method has also been used to isolate exon sequences of a specific gene (i.e., the DNA repair gene, ERCC1) from randomly selected genomic clones known to be derived from a segment of human chromosome 19. The sensitivity and ease of the exon amplification

30 method is such that 20-40 kbp of genomic DNA have been screened in a single transfection. The subject method can be used for rapid identification of

-5-

transcribed segments of mammalian genomes and the generation of chromosomal transcription maps.

#### Brief Description of the Drawings

Figure 1 is a schematic representation of the structure of the in vivo splicing plasmid pSPL1 and a schematic representation of the exon amplification method of the present method.

Figure 2 shows the results of sequence analysis of amplified product derived from  $\lambda$  genomic clone 5W.

#### Detailed Description of the Invention

The present invention relates to a method of isolating exon sequences from genomic DNA in which exon sequences are rescued by virtue of selection for functional 5' and 3' splice sites. The method of the present invention, referred to as exon amplification, is useful for isolating coding sequences from complex mammalian genomic DNA and for screening 20-40 kbp of genomic DNA in a single transfection. It is based upon the ability of RNA transcripts to undergo processing events which require the presence of splice recognition and branchpoint sequences. In the present method, random segments of chromosomal DNA are inserted into an intron present within an in vivo splicing plasmid, the resulting in vivo splicing plasmid construct is transfected into an appropriate mammalian host cell, and the transfected host cell is maintained under conditions appropriate for replication of the in vivo splicing plasmid

-6-

construct. Transcription from the promoter of the splicing construct, followed by RNA processing results in production of cytoplasmic mRNA, which is screened by PCR amplification for the acquisition of an exon from the genomic fragment. In particular, as described herein, fragments of cloned genomic DNA are inserted into the intron contained within an in vivo splicing plasmid and the resulting construct is introduced into an appropriate host cell (one in which the in vivo splicing plasmid construct can replicate or be amplified). If the fragment introduced into the vector contains an exon, the vector splice sites are paired with splice sites of the inserted fragment. The resulting mature cytoplasmic RNA contains the previously unidentified exons, which are cloned by known techniques, for example by means of RNA-based PCR amplification and cloning.

Previous studies have shown that introns constructed with novel combinations of 5' and 3' splice sites from diverse genes are actively spliced. Thus, this method is generally applicable for the selection of exon sequences from any gene. The method is also rapid and easily adapted to large scale experiments. A series of cloned genomic DNA fragments can be screened within one to two weeks. The sensitivity of this method is high. Genomic DNA segments of 20 kb or more can be successfully screened in a single transfection using a set of pooled subclones. This method thus allows the rapid identification of exons in mammalian genomic DNA and should facilitate the isolation of a wide spectrum

-7-

of genes of significance in physiology and development.

In the present method, genomic DNA from any mammalian cell type can be analyzed for the presence  
05 of exons. DNA to be analyzed is obtained (e.g., from a cell or cloned genomic DNA, such as a cosmid library) and fragmented, using known methods and introduced into an appropriate in vivo splicing plasmid, such as the pSPL1 vector described herein.  
10 DNA to be introduced into the in vivo splicing plasmid is digested with appropriately-selected restriction enzyme(s), (i.e., restriction enzyme(s) which make it possible for the resulting fragments to be introduced into an insertion or cloning site  
15 or sites in a plasmid intron in the in vivo splicing plasmid). For example, in the case of pSPL1, DNA is digested with BamHI, resulting in DNA fragments which can be inserted into the unique BamHI site in the plasmid intron (see Figure 1). The resulting  
20 fragments are subcloned into the in vivo splicing plasmid at the insertion site and the resulting in vivo splicing plasmid construct is introduced (e.g., using transfection by electroporation) into an appropriate host cell, in which replication or  
25 amplification of the plasmid occurs. RNA containing the introduced genomic fragment is produced and is processed in such a way that the mature RNA (mRNA) includes the exon.

The pSPL1 vector is representative of in vivo  
30 splicing plasmids useful in the present method of exon amplification. Other plasmid vectors which contain a promoter, such as the SV40 or CMV

-8-

promoter, whose presence and function in the in vivo splicing plasmid results in high levels of transcription, and which contain a plasmid intron into which a DNA fragment can be inserted, including  
05 splice junction or splice sites, and a branchpoint recognition sequence can be used. Any convenient insertion site or sites present in the intron and located 5' to the branchpoint recognition sequence will be useful for efficient splicing of the exon  
10 introduced into the splicing plasmid. In one embodiment, the splicing vector can be engineered using known techniques to introduce a multiple cloning site into the splicing vector's intron, preferably 5' of the branchpoint sequence, to  
15 increase the available convenient restriction sites. Particularly useful are vectors which include an origin of replication under the control of which amplification occurs in the host cells, increasing the number of templates for transcription. A  
20 bacterial origin of replication and selectable marker are useful for propagation of the plasmids in bacterial hosts, and a mammalian polyadenylation signal is useful for message stability, efficient processing and transport of transcripts in the  
25 mammalian host cell.

Splice junctions or splice sites, such as those of the HIV-1 tat gene present in pSLP1, which are not as efficient as others, are also particularly useful. The 5' and 3' splice site sequences  
30 flanking the pSPL1 vector exons, which were selected to minimize exon skipping, are a particular advantage of the pSPL1 vector. These splice sites,

-9-

derived from the tat exons of HIV-1, are slowly spliced in both in vivo and in vitro systems. The inefficient splice sites of tat are therefore compatible for reactions with splice sites from  
05 unrelated genes, and have been shown to be efficiently spliced to sites flanking the exons of the rat preproinsulin and the rabbit  $\beta$ -globin genes.

The transcribed sequences flanking the intron in the splicing vector are preferably of known  
10 sequence to facilitate selection of appropriate primers for amplification. In addition, these sequences preferably do not contain repetitive sequences, termination signals or additional functional splicing signals, which may interfere  
15 with joining of the genomic exon sequences with the desired 5' and 3' splice sites flanking the splicing vector's intron.

In a further embodiment, the splicing vector can be modified such that the products resulting  
20 from self-splicing (i.e. joining of the vector's 5' and 3' splice sites together, rather than to genomic sequences) can be minimized prior to amplification by destruction of those templates. For example, the region immediately adjacent to the splice  
25 recognition sites can be altered such that, when the vector sequences are spliced together instead of to a genomic sequence (e.g. when there is no insert, or when the genomic exon is skipped), a restriction endonuclease site is created in the cDNA product of  
30 the transcript. Following the synthesis of double stranded cDNA from isolated mature RNA and prior to amplification, digestion with the appropriate

-10-

restriction enzyme cleaves these templates. Because the cleavage site is between the primer binding sites, the double stranded amplification product from these templates is minimized or abolished. For  
05 example, a restriction enzyme, such as BstXI, which recognizes an interrupted palindrome can be used. This enzyme recognizes the sequences CCA and TGG separated by six nucleotides of unspecified sequence, thus, the splice recognition sequences can  
10 be accommodated in the unspecified nucleotides. This process can result in an enrichment of products containing a mammalian exon, and increases the sensitivity of the procedure in detecting such exons. Thus, the amount of genomic DNA that can be  
15 screened for exons in one step is increased substantially. For example, YAC clones with inserts some 200-500 kb can be accommodated by increasing the sensitivity in this manner.

Cytoplasmic RNA produced in host cells  
20 containing the in vivo plasmid construct is isolated from the host cells and subjected to amplification (e.g., by RNA-based PCR, as described in Example 1). Following the reverse transcription and amplification reactions resulting in DNA production, the  
25 RNA/PCR product is cloned by purifying the appropriate DNA product, digesting it with appropriate restriction enzymes (e.g., SalI and MscI) and cloning it into appropriate restriction sites (e.g., SalI and EcoRV sites, respectively) in an appropriate  
30 plasmid (e.g., pBluescript/11SK<sup>+</sup>, Stratagene). Cloned products are then sequenced, using known techniques, such as the dideoxy chain termination

-11-

method. Purification of the PCR product is not necessary prior to cloning, when a suitable probe for identification of constructs containing genomic exons is available. The PCR products can also be  
05 sequenced directly without cloning. For example, in one convenient approach, the appropriate PCR fragment can be purified, reamplified with primers adjacent to the vector splice sites (and, therefore, adjacent to the genomic exon) and sequenced.

10 The method of exon amplification described herein is a rapid and efficient technique for the identification of expressed DNA sequences in complex mammalian genomes. This method circumvents the laborious characterization of a cloned genomic DNA  
15 segment and permits a direct transition to a cDNA. The initial need for appropriate sources of RNA for isolation of cDNA clones is thus also circumvented. The efficacy of exon amplification has been clearly demonstrated, as described particularly in the  
20 Examples, by the identification and cloning of exons from a cosmid known to contain a portion of the mouse Na,K-ATPase  $\alpha 1$  subunit gene, as well as exon sequences of the human DNA repair gene, ERCC1, from an uncharacterized  $\lambda$  genomic clone. The present  
25 method of exon amplification can be used, for example, for rapidly determining the tissues in which a particular gene is expressed, either by northern analysis or by in situ hybridization. It can also be of use in the isolation of complete  
30 cDNAs by library screening procedures or by anchored-PCR techniques (Loh, E.Y., et al., Science 243:217-220 (1989)).

-12-

Exon amplification can also be used to complement recently developed methods for isolating transcribed segments of the human genome. Two such methods involve the use of human-hamster hybrid cell lines containing specific regions of the human genome. The first approach involves the generation of a cDNA library from heteronuclear (hn) RNA using oligonucleotides complementary to consensus 5' intron splice sequences (Duyk, G.M., et al., Proc. Natl. Acad. Sci. USA 87:8995-8999 (1990), followed by screening for human-specific repetitive sequences. The second method utilizes oligonucleotides from the conserved region of Alu repetitive sequences to generate cDNAs from hnRNA by PCR. (Corbo, L. et al., Science 249: 652-655 (1990)). The efficacy of both of these strategies is, however, dependent on significant expression of human-specific RNA in the hybrid cells. The products derived from these approaches must also be made free of repetitive sequences prior to use as probes for screening or blotting experiments, due to the presence of repetitive sequences in the introns of these unspliced RNAs. This problem can be resolved by the application of exon amplification, using cDNAs from hnRNA as starting material. Since these cDNAs represent cloned transcription units, the combination of these approaches should greatly facilitate the cloning of coding sequences.

Use of the present method of exon amplification in large scale screening for transcribed sequences may provide a new approach to genetic mapping. For instance, the construction of transcription maps for

-13-

large segments of mammalian genomes is technically feasible using this method. Such an approach could provide a powerful adjunct in the fine mapping of the human genome, and would enhance the efficiency with which genes responsible for numerous genetic disorders are identified.

The nature of the sequence and structure specificity underlying the selection of exons during the splicing of normal nuclear precursor RNAs is not well understood. This specificity is sufficient to screen introns over 100,000 nucleotides in length in the accurate joining of the flanking exons. Experiments suggest that this remarkable specificity is not dictated by the unique nature of the two exons flanking an intron. In fact, all 5' and 3' splice sites are thought to be generically compatible for accurate splicing. This is typified by the accurate splicing of a hybrid intron where the 5' splice site is derived from an exon of the rat preproinsulin gene and the 3' splice site is from a viral exon. These results suggest that the present exon amplification method should be able to identify most of the exons within a genomic fragment.

The present method will now be illustrated by the following examples, which are not intended to be limiting in any way.

EXAMPLE 1 Isolation and Cloning of Exon Sequences from Individual Segments of Genomic DNA

Fragments of a mouse cosmid clone, MaG#9, known to contain exon sequences of the Na, K-ATPase  $\alpha 1$  subunit gene (Tam, S.-Y., et al., Mol. Cell. Biol.

-14-

10:6619-6623 (1990) were used as described below to demonstrate the ability of the present method to amplify exons from complex mammalian genomic DNA. Below is a description of the materials and methods used throughout the work described herein, followed by a description of their use with the mouse clone to identify the specific subunit gene.

#### Cell Culture and Electroporation

COS7 cells (clonal line A6) were propagated in DME medium supplemented with 10% inactivated fetal calf serum. For transfections COS7 cells were grown to 75-85% confluency, trypsinized, collected by centrifugation, and washed in ice-cold phosphate buffered saline (PBS) in the absence of divalent cations. The washed cells (approximately  $4 \times 10^6$ ) were then resuspended in cold 0.7 ml PBS and combined in a precooled electroporation cuvette (0.4 cm chamber, Bio Rad) with 0.1 ml PBS containing 1-15  $\mu$ g DNA. After 10 minutes on ice the cells were gently resuspended, electroporated (1.2 kV[3 kV/cm], 25  $\mu$ f) in a Bio Rad Gene Pulser, and placed on ice again. After 10 minutes the cells were transferred to a tissue culture dish (100 mm) containing 10 ml prewarmed, preequilibrated culture medium.

#### 25 Vector Construction and Oligonucleotides

pSPL1 was constructed as follows: A 2.7 kbp TagI fragment from pgTat (corresponding to nucleotides 68-2775 of HIV isolate HXB3) (Malim, M.M. et al., Nature (London) 335:181-183 (1988) was cloned into a SalI site of pBluescript+ (Stratagene). A

-15-

2.6 kbp BamHI-PstI fragment was isolated from this construct and used to replace the BamHI-EcoRI region of pS $\beta$ -IVS2 (Buchman, A. et al., Mol. Cell. Biol. 8:4395-4405 (1988), a shuttle vector containing the  
05 SV40 origin of replication and early region promoter upstream of rabbit  $\beta$ -globin sequences, including  $\beta$ -globin IVS2. This results in removal of  $\beta$ -globin IVS2 and addition of HIV-tat intron and flanking exon sequences. The EcoRI and PstI sites were  
10 removed by blunt-end cloning. The BamHI site in this construct was subsequently removed by BamHI digestion followed by blunting with mung bean nuclease. Finally, a BamHI site was inserted into the HIV-tat intron at the unique KpnI site.  
15 Oligonucleotide pairs and the predicted lengths of the PCR products generated by spliced RNA from the vector are: DHA15,CCAGTGAGGAGAAGTCTGCGG and DHAB14,GTGAGCCAGGGCATTGGCC:689 bp product; SD2,GTGAACTGCACTGTGACAAGC and  
20 SA2,ATCTCAGTGGTATTTGTGAGC: 429 bp product; SD1,CCCGGATCCGCGCGACGAAGACCTCCTCAAGGC (BamHI cloning site at 5'-end) and SA1,CCCGTCGACGTCGGGTCCCCTCGGGATTGG (Sall cloning site 5'-end): 102 bp product. The antisense oligo-  
25 nucleotide (DHAB14 and SA2) were used as primer in the first strand cDNA synthesis reactions (see below). SD1 and SA1 are internal to the initial RNA/PCR product and were used for reamplification of RNA/PCR products.

-16-

RNA Isolation, RNA/PCR Amplifications, and Cloning

Cytoplasmic RNA was isolated 48-72 hours post-transfection. Briefly, cells were washed three times with ice cold PBS, scraped in 10 ml ice cold  
05 PBS, and collected by centrifugation. The cell pellet was then resuspended on ice in a low ionic strength buffer (10 mM Tris pH 7.5, 1 mM KCl, 1 mM MgCl<sub>2</sub>) and cells were lysed by the addition of 0.5% (cf) Triton X-100. Nuclei were removed by centri-  
10 fugeation, 0.5% SDS (cf) was then added to the supernatant, which was subsequently extracted with Tris-buffered phenol followed by phenol:chloroform (1:1). RNA was precipitated by the addition of 0.2 M NaCl (cf) and 2.5 volumes of ethanol, followed by  
15 storage at -20°C. RNA was quantitated by A<sub>260</sub> determination.

First strand cDNA synthesis was performed as follows: RNA (2.5 or 5 µg) was added to a reverse transcription solution consisting of 10 mM Tris pH  
20 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.001% gelatin, and the mixture was heated to 65°C for 5 minutes. 3.5 U RNasin (Promega) and 200 U MMLV reverse transcriptase (BRL) were added to the reaction (final volume: 25 µl), which was then incubated at 42°C for 90-120  
25 minutes.

The entire reverse transcription reaction was then subjected to PCR amplification in a Thermocycler (Perkin-Elmer-Cetus) using the appropriate oligonucleotide pairs. Thirty-five amplification  
30 cycles were routinely used, and consisted of 1 minute at 94°C, 2 minutes at 55-58°C and 3 minutes at 72°C. Products were visualized by staining with

-17-

ethidium bromide after electrophoresis in 1-1.5% agarose gels.

To clone the RNA/PCR product, the appropriate DNA fragment was purified from low melting point agarose and digested with SalI and MscI, unique restriction sites present within the vector ( $\beta$ -globin) sequences, and cloned into the SalI and EcoRV sites of pBluescript11SK+ (Stratagene). Alternatively, the gel-purified product was subjected to a second PCR amplification using the internal oligonucleotide pair SD1 and SA1, which flank the vector splice junctions and contain BamHI and SalI cloning sites, respectively. The product from this reaction was gel purified, end-repaired with T4 DNA polymerase (New England Biolabs), digested with BamHI and SalI, and cloned into pBluescript11SK+. Cloned products were sequenced using the dideoxy chain termination method (Sanger, F. et al., Proc. Natl. Acad. Sci. USA 74:5463-5467 (1977)).

#### Blot analysis

Restriction endonuclease digested genomic DNA clones were electrophoresed through 0.8% or 0.9% agarose gels. RNA/PCR products were electrophoresed through 1-1.5% agarose gels. RNA samples were electrophoresed through a 1% agarose, 6% formaldehyde gel and blotted onto a GeneScreen Plus membrane (New England Nuclear). Filters were hybridized using standard procedures. Sambrook, J. et al., Molecular Cloning: A Laboratory Manual (2nd ed.), Cold Spring Harbor Laboratory Press (1989). DNA

-18-

probes were radiolabelled to high specific activity using the random primer method (Feinberg, A.P. et al., Biochem. Biophys. Res. Commun. 111:47-54 (1983)).

05       The strategy for exon amplification is outlined in Figure 1. A vector, designated pSPL1, was designed to make it possible to insert mammalian genomic DNA segments approximately 1 to 4 kbp long at a BamHI site present within an intron in the  
10 vector. Construction of this vector is described in detail in Example 1. The insertion site is within an intron from the HIV-1 tat gene whose flanking exons and splice sites were substituted for the second intron of the rabbit  $\beta$ -globin gene. The  
15 reporter gene is transcribed by the SV40 early promoter and a polyadenylation signal is derived from SV40. Upon transfection of this plasmid construct into COS7 cells, RNA transcripts are efficiently generated and the tat intron sequences  
20 are spliced to produce a polyadenylated cytoplasmic RNA.

When a fragment containing an entire exon with flanking intron sequence (in the proper orientation) is inserted into the BamHI site of the pSPL1 vector,  
25 the exon should be retained in the mature poly A+ cytoplasmic RNA.

#### Isolation and Cloning of Exon Sequences of the Na,K-ATPase $\alpha$ 1 subunit gene

A known source (fragments of a mouse cosmid  
30 clone, MoG#9) of exon sequences of a gene (the Na,K-ATPase  $\alpha$ 1 subunit gene (Tam, S.-Y. et al., Mol.

-19-

Cell. Biol. 10:6619-6623 (1990)) was used to demonstrate that when a fragment containing an entire exon with flanking intron sequence (in the proper orientation) is inserted into an intron of an in vivo splicing plasmid, the exon is retained in the mature poly A<sup>+</sup> cytoplasmic RNA. Fragments of the mouse cosmid clone were subcloned into pSPL1. A 3.5 kbp BglIII fragment of the cosmid was subcloned into the BamHI site of pSPL1 in sense and antisense orientations and the resulting constructs were introduced by transfection into COS7 cells. Cytoplasmic RNA preparations derived from the transfectants were analyzed by northern blotting, using a radiolabeled NcoI-BamHI fragment spanning nucleotides 111-705 of the murine Na, K-ATPase  $\alpha$ 1 cDNA as probe (Kent, R.B., et al., Science 237:901-903 (1987)). An abundant 2.2 kb RNA species was detected only in cells transfected with the sense construct, indicating expression and processing of the transfected sequences.

To isolate spliced exons contained within the vector-derived RNA sequences, an RNA-based PCR (RNA/PCR) method was used. Cosmid M $\alpha$ G#9 was digested with either BamHI or BglIII, or with the combination of these endonucleases, followed by "shotgun" cloning into pSPL1. These constructs were then screened for the presence of exon sequences as described above, using  $\beta$ -globin specific oligodeoxynucleotides (oligonucleotides SD2 and SA2) as RNA/PCR primers. The sense and antisense constructs described in Figure 1a were similarly analyzed. The resulting RNA/PCR products were visualized by

-20-

electrophoresis through 1.5% agarose gels and staining with ethidium bromide. As expected, oligodeoxynucleotide primers SD2 and SA2 generated an RNA/PCR product of 429 bp from RNA of transfectants with the pSPL1 vector. The product migrating at 429 bp is derived from splicing occurring between vector 5' and 3' splice sites. A ~300 bp product is present in all lanes, including mock-transfected (no DNA) cells, indicating that this product is an artifact derived from the COS7 cell background. Analysis of RNA from COS7 cells transfected with the 3.5 kbp BglIII fragment inserted in the sense orientation into pSPL1 yielded a PCR product of approximately 1.5-1.6 kbp. Transfection of a recombinant containing the same fragment in the opposite orientation only yielded the 429 bp PCR product containing vector sequences. Hybridization of the mouse  $\alpha 1$  subunit cDNA to blots containing these RNA/PCR products confirmed that sequences derived from the sense construct consist of ATPase exons. "Sense" and "antisense" RNA/PCR products from an experiment similar to that described above were blotted and hybridized to the NcoI-BamHI fragment probe. The larger size of the product detected in the "sense" lane (approximately 1.8 kbp), when compared to the other product generated, is due to use of the oligonucleotide pair DHAB14 and DHAB15 in the RNA/PCR reaction, which will amplify 689 bp of vector sequence. The length and restriction pattern of the RNA/PCR product derived from sense transfectants are consistent with proper splicing of six ATPase exons.

-21-

A more detailed analysis on the RNA/PCR product generated by a 2.8 kbp BglII fragment from the MαG#9 cosmid was also performed. Insertion of this fragment into pSPL1 and transfection yielded a 600  
05 bp RNA/PCR product, which was subsequently cloned and sequenced, using the methods described above. This product contained exon sequences of the α1 cDNA, spanning 171 bp from bp 125-295. This represents precisely two exons of the gene, whose  
10 sequence and structure has recently been characterized. This is proof that accurate processing occurred between tat and α1 splice recognition sequences, resulting in the removal of the HIV-tat and ATPase intron sequences, and the insertion of  
15 ATPase exons in the vector-derived mature RNA.

EXAMPLE 2 Isolation and Cloning of Multiple Genomic DNA Fragments

Thus, as described in Example 1, it has been shown that, in its simplest form, the in vivo  
20 splicing selection method of the present invention can be used to amplify exon sequences from individual segments of genomic DNA. However, in situations in which large regions of a chromosome require analysis in this manner, examination of  
25 single fragments can be extremely cumbersome. Therefore, assessment of whether multiple fragments could be analyzed simultaneously was also carried out, as described in this example. The materials and methods used were as described in Example 1.  
30 The Na,K-ATPase α1 subunit cosmid, MαG#9, was digested separately with BamHI, BglII, or with the

-22-

combination of BamHI plus BglII. Each digest was subsequently "shotgun" cloned into pSPL1. These mixtures of clones were then transfected into COS7 cells and the resulting RNA was analyzed by RNA/PCR  
05 for the presence of products larger than that from pSPL1 alone. In this situation the predominant RNA will contain only sequences from the vector pSPL1, since the majority of genomic fragments contain no exon sequences or are inserted in the antisense  
10 orientation. PCR analysis of RNA preparations from cells transfected with "shotgun" clones of BamHI, BglII, or BamHI plus BglII digestions of MaG#9 generated multiple products larger than the 429 bp derived from pSPL1. The BglII RNA/PCR product was  
15 gel purified, radiolabelled and directly hybridized to a BglII restriction digest of MaG#9. Hybridization of this product to the 2.8 kbp BglII genomic fragment demonstrated that the amplified product was derived from a genomic fragment known to contain an  
20 exon. These results indicate that in a situation where the complexity of the genomic DNA is high, exon sequences can still be identified in a single transfection. The 1.6 kbp product detected following transfection with the 3.5 kbp BglII sense  
25 construct was not observed in the BglII "shotgun" transfection RNA/PCR product(s). This is most likely due to competition among PCR templates, favoring smaller and more abundant substrates. Also, a weakly staining product migrating at  
30 approximately 650 bp was observed in nearly all reactions containing RNA from plasmid (including

-23-

pSPL1 alone) transfections and are likely to be artificial.

EXAMPLE 3 Screening of Complex Human Genomic DNA

To further test the ability of the exon  
05 amplification method to screen complex genomic DNA  
for the presence of exons, genomic clones containing  
15-20 kbp human genomic DNA inserts were analyzed.  
Twelve previously uncharacterized human genomic  $\lambda$   
phage clones, derived from a radiation-reduced  
10 human-hamster hybrid cell line containing a segment  
of human chromosome 19, were digested with BamHI  
plus BglII, "shotgun" cloned into pSPL1, and trans-  
fected into COS7 cells. RNA preparations from these  
transfectants were examined by RNA/PCR. Six of the  
15 12 amplification reactions designated (1B, 5B, 5C,  
5W, 6B and 6C) clearly generated products larger  
than the vector-derived 429 bp product, suggesting  
that exon sequences are present in each of these  
clones. The products from 1B (600-620 bp doublet),  
20 5C (600 bp product) and 5W (620 bp product) were  
excised from agarose gels,  $^{32}$ P-labelled using the  
random primer method (Feinberg, A.P. et al.,  
Biochem. Biophys. Res. Commun. 111:47-54 (1983)),  
and hybridized to filters containing blotted DNAs  
25 from the original genomic clones. Each product  
hybridized only to the genomic DNA segment from  
which it was derived, indicating that the amplified  
sequences were not derived from  $\lambda$  phage DNA. The  
absence of cross-hybridization to other human DNA  
30 fragments indicated that the PCR products were  
essentially free of repetitive sequences. In some

-24-

cases, two genomic fragments were detected by these probes, suggesting that more than one PCR product was present.

Four of these PCR products were reamplified and  
05 cloned using internal oligonucleotides (SD1 and SA1)  
which correspond to sequences immediately flanking  
the plasmid splice donor and acceptor sites, and  
which contain artificial cloning sites. This was  
followed by cloning into pBluescriptIISK+. Sequence  
10 analysis of clones from one of these products,  
derived from phage 5W, revealed that the RNA/PCR  
product was derived from an exon of the DNA excision  
repair gene ERCC1. This gene is located on human  
chromosome 19 and is known to be present in the  
15 human-hamster hybrid cell line from which the  
genomic clones were derived. A perfect match of the  
sequence between the HIV tat splice junctions and  
bases 136-247 of the ERCC1 cDNA sequence (van Duin,  
M. et al., Cell 44:913-923 (1986)) indicates that an  
20 exon of this gene has been rescued.

#### EXAMPLE 4 Screening of Uncharacterized Regions of the Human Genome

The present method of exon amplification has  
also been extended to uncharacterized regions of the  
25 human genome. In preliminary studies, approximately  
70% of cosmid genomic clones (23/33) and 45% of  $\lambda$   
phage genomic clones (8/18) have yielded RNA/PCR  
products containing potential exon sequences. cDNAs  
corresponding to 6 of these products are currently  
30 under characterization. These results demonstrate  
the effectiveness of the exon amplification method

-25-

in the identification of exon sequences in otherwise uncharacterized genomic DNA clones.

-26-

CLAIMS

1. A method of isolating a coding sequence from mammalian genomic DNA, comprising the steps of:

- 05 a) providing fragmented mammalian genomic DNA;
- b) inserting fragmented mammalian genomic DNA into an intron of an in vivo splicing plasmid, thereby producing an in vivo splicing plasmid construct having
- 10 mammalian genomic DNA within a plasmid intron;
- c) introducing the in vivo splicing plasmid construct into a host cell in which it replicates;
- 15 d) maintaining the product of (c) under conditions appropriate for replication of the in vivo splicing plasmid construct and transcription of DNA present in the in vivo splicing plasmid construct, thereby
- 20 producing a host cell containing mature RNA; and
- e) isolating mature RNA produced in (d).

2. The method of Claim 1 further comprising the steps of:

- 25 f) amplifying the product obtained in (e) using an RNA-based amplification method, thereby producing cDNA of the mammalian genomic DNA;
- g) digesting the product of (f) with a
- 30 restriction enzyme, which recognizes a

-27-

restriction site present in a cloning vector; and

- 05 h) cloning the product of (g) in the cloning vector, thereby producing cloned DNA of the mammalian genomic DNA.

3. The method of Claim 2 wherein step (f) further comprises purifying the cDNA of the mammalian genomic DNA.

- 10 4. The method of Claim 2 further comprising determining the nucleotide sequence of the cloned cDNA of the mammalian genomic DNA.

5. The method of Claim 1 further comprising the step of:

- 15 f) screening mature RNA obtained in (e) with appropriate antisense probes, thereby identifying mature RNA containing a mammalian genomic DNA exon.

6. The method of Claim 5 further comprising the steps of:

- 20 g) amplifying the product obtained in (e) using an RNA-based amplification method, thereby producing cDNA of the mammalian genomic DNA;
- 25 h) digesting the product of (g) with one or more restriction enzymes, which recognize a restriction site present in a cloning vector; and

-28-

- i) cloning the product of (h) in the cloning vector, thereby producing cloned DNA of the mammalian genomic DNA.
- 05 7. The method of Claim 1 further comprising the steps of:
- f) amplifying the product obtained in (e) using an RNA-based amplification method, thereby producing cDNA of the mammalian genomic DNA;
  - 10 g) purifying cDNA of the mammalian genomic DNA; and
  - h) sequencing the cDNA obtained in (g).
- 15 8. The method of Claim 7, wherein step (f) further comprises purifying the cDNA and reamplifying with PCR primers comprising sequences complementary to sequences adjacent to the splice sites of the splicing vector.
- 20 9. A method of exon amplification, comprising the steps of:
- a) providing fragmented mammalian genomic DNA;
  - b) inserting fragmented mammalian genomic DNA into an intron of an in vivo splicing plasmid in which the splice sites present are inefficient or slowly spliced, thereby
  - 25 producing an in vivo splicing plasmid construct having mammalian genomic DNA within a plasmid intron;

-29-

- c) introducing the splicing plasmid construct of (b) into a host cell in which it replicates;
  - 05 d) maintaining the product of (c) under conditions appropriate for replication of the in vivo splicing plasmid construct and transcription of DNA present in the in vivo splicing plasmid construct, thereby producing a host cell containing mature RNA;
  - 10 e) obtaining mature RNA which includes a mammalian genomic exon;
  - f) producing DNA by reverse transcription of the mature RNA obtained in (e); and
  - 15 g) amplifying DNA produced in (f).
10. The method of Claim 9 further comprising sequencing the product of step (g).
11. The method of Claim 9 wherein the mammalian genomic DNA is selected from the group consisting of: human DNA and mouse DNA.
- 20
12. A method of exon amplification, comprising the steps of:
- a) providing fragmented mammalian genomic DNA;
  - 25 b) inserting fragmented mammalian genomic DNA into the in vivo splicing plasmid pSPL1 at a restriction site in the HIV-tat intron present in pSPL1, thereby producing a

-30-

- pSPL1 construct having mammalian genomic DNA present with the HIV-tat intron;
- 05 c) introducing the pSPL1 construct into a host cell in which it replicates;
- d) maintaining the product of (c) under conditions appropriate for replication of the pSPLI construct and transcription of DNA present in the pSPLI construct, thereby producing a host cell containing
- 10 mature RNA;
- e) obtaining mature RNA which includes a mammalian genomic exon;
- f) producing DNA by reverse transcription of the mature RNA obtained in (e); and
- 15 g) amplifying DNA produced in (f).
13. The method of Claim 12 further comprising sequencing the product of step (g).
14. The method of Claim 12 wherein the mammalian genomic DNA is selected from the group
- 20 consisting of: human DNA and mouse DNA.
15. The method of Claim 14 wherein the host cell in (c) is a COS7 cell.
16. An in vivo splicing plasmid comprising:
- 25 a) an origin of replication that is functional in an appropriate host cell;
- b) a promoter sequence;

-31-

- 05 c) an intron and transcribed flanking sequences under the control of said promoter, including therein a 5' splice site, a 3' splice site, a branchpoint recognition sequence, and a polyadenylation signal, the intron including one or more cloning sites 5' to the branchpoint recognition sequence;
- 10 d) a bacterial origin of replication; and  
e) a selectable marker.
17. The in vivo splicing plasmid of Claim 16 wherein the 5' and 3' splice sites allow joining with sites inserted into the intron of the splicing plasmid.
- 15 18. An in vivo splicing plasmid comprising:
- a) an SV40 origin of replication and early region promoter; and
- 20 b) a  $\beta$ -globin gene in which the second intron has been replaced by an HIV-tat gene intron and flanking exon sequences, including therein the HIV-tat gene 5' splice site, 3' splice site, and branchpoint recognition sequence, the HIV-tat gene intron including a cloning
- 25 site.
19. The in vivo splicing plasmid pSPL1.
20. A kit comprising:

-32-

- a) an in vivo splicing plasmid comprising:
- 1) an origin of replication that is functional in an appropriate host cell;
  - 05 2) a promoter sequence;
  - 3) an intron and transcribed flanking sequence under the control of said promoter, including therein a 5' splice site, a 3' splice site, a  
10 branchpoint recognition sequence, and a polyadenylation signal, the intron including one or more cloning sites 5' to the branchpoint recognition sequence;
  - 15 4) a bacterial origin of replication; and
  - 5) a selectable marker; and
- b) appropriate oligonucleotide primers which hybridize to transcribed sequences  
20 flanking said intron.

21. The kit of Claim 20, wherein the 5' and 3' splice sites of the in vivo splicing plasmid allow joining with sites inserted into the intron of the splicing plasmid.

25 22. A kit comprising:

- a) an in vivo splicing plasmid comprising:
- 1) an SV40 origin of replication and early region promoter; and

-33-

- 2) a  $\beta$ -globin gene in which the second intron has been replaced by an HIV-tat intron and flanking exon sequences, including therein the HIV-tat gene 5' splice site, 3' splice site, and branchpoint recognition sequence, the HIV-tat gene intron including a cloning site; and
- b) appropriate oligonucleotide primers which hybridize to the  $\beta$ -globin exon sequences flanking said HIV-tat intron.
23. A kit comprising:
- a) the in vivo splicing plasmid pSPL1; and
- b) oligonucleotide primers which hybridize to transcribed sequences flanking the HIV-tat intron within said plasmid.
24. The kit of Claim 23, wherein the oligonucleotide primers comprise one or more of the following:
- a) DHA15, CCAGTGAGGAGAAGTCTGCGG and DHAB14, GTGAGCCAGGGCATTGGCC;
- b) SD2, GTGAAGTGCCTGTGACAAGC and SA2, ATCTCAGTGGTATTTGTCAGC;
- c) SD1, CCCGGATCCGCGCGACGAAGACCTCCTCAAGGC and SA1, CCCGTCGACGTCGGGTCCCTCGGGATTGG.

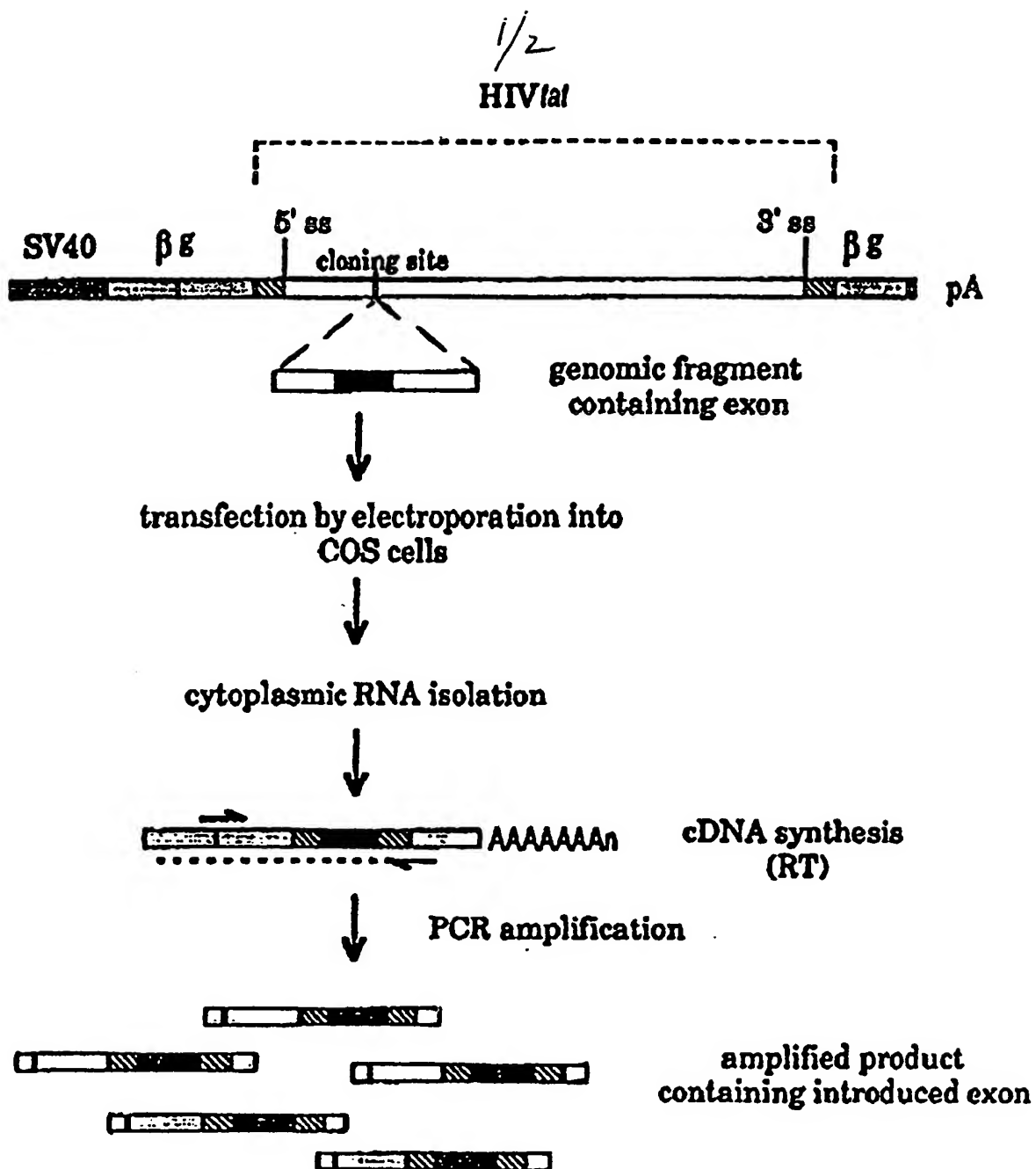
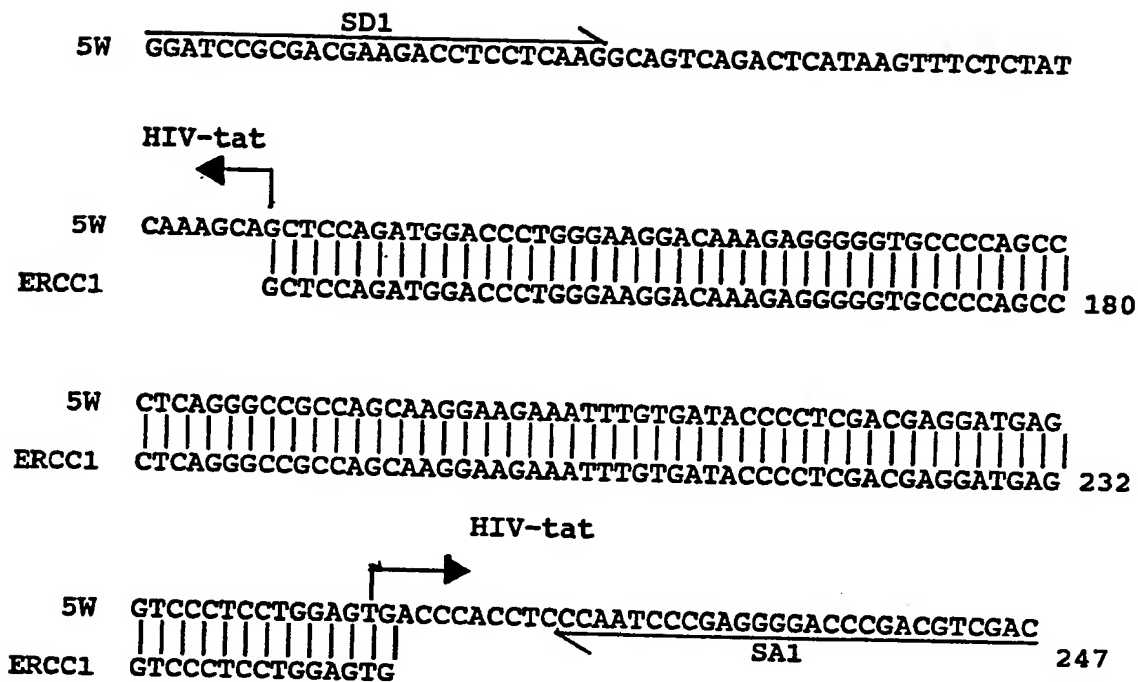


Figure 1

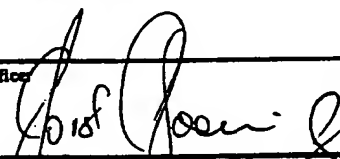
$\frac{2}{2}$ 

**Figure 2**

## INTERNATIONAL SEARCH REPORT

PCT/US 92/00692

International Application No

<b>I. CLASSIFICATION</b> <b>F SUBJECT MATTER</b> (If several classification symbols apply, indicate all) <sup>6</sup>		
According to International Patent Classification (IPC) or to both National Classification and IPC		
Int.Cl. 5 C12N15/00; C12N5/10;	C12N15/67; C12Q1/68	C12N15/85; C12N15/11
<b>II. FIELDS SEARCHED</b>		
Minimum Documentation Searched <sup>7</sup>		
Classification System	Classification Symbols	
Int.Cl. 5	C12N	
Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched <sup>8</sup>		
<b>III. DOCUMENTS CONSIDERED TO BE RELEVANT<sup>9</sup></b>		
Category <sup>10</sup>	Citation of Document, <sup>11</sup> with indication, where appropriate, of the relevant passages <sup>12</sup>	Relevant to Claims No. <sup>13</sup>
X	NUCL. ACID RES. vol. 18, no. 22, 25 November 1990, IRL PRESS, OXFORD, ENGLAND; pages 6743 - 6744; D. AUCH AND M. RETH: 'Exon trap cloning: using PCR to rapidly detect and clone exons from geneomic DNA fragments'	1-11
Y	see figures 1,2	12-15, 18-24
X	EMBO JOURNAL vol. 2, no. 5, May 1983, IRL PRESS LIM., OXFORD, ENGL.; pages 727 - 733; T. KUHNE ET AL.: 'Evidence against a scanning model of RNA splicing'	16,17
Y	see figures 1,2	18,19
<p><sup>10</sup> Special categories of cited documents:</p> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> <p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"A" document member of the same patent family</p>		
<b>IV. CERTIFICATION</b>		
Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report	
21 MAY 1992	- 9. 06. 92	
International Searching Authority	Signature of Authorized Officer	
EUROPEAN PATENT OFFICE	HORNIG H. 	

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)		
Category *	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claims No.
X	CELL vol. 37, July 1984, CELL PRESS, CAMBRIDGE, NA.; pages 915 - 925;	16,17
Y	B. WIERINGA ET AL.: 'A minimal intron length but no specific internal sequence is required for splicing the large rabbit beta-globin intron' see page 915, right column, line 48 - line 57; figure 7 ---	18,19
Y	NATURE vol. 335, 8 September 1988, MACMILLAN JOURNALS LTD., LONDON, UK; pages 181 - 183; M.M. MALIM ET AL.: 'Immunodeficiency virus rev trans-activator modulates the expression of the viral regulatory sequences' cited in the application ---	12-15, 18-24
P,X	PROC. NATL. ACAD SCI. vol. 88, no. 9, 1 May 1991, NATL. ACAD SCI., WASHINGTON, DC, US; pages 4005 - 4009; A.J. BUCKLER ET AL.: 'Exon amplification: A strategy to isolate mammalian genes based on RNA splicing' ---	1-24
P,X	AM. J. HUM. GENET. (4 SUPPL.) vol. 49, 1991, UNIV. CHICAGO PRESS, CHICAGO, US; page 420; S.A.M. TAYLOR ET AL.: 'The isolation of a cDNA from the huntington's disease candidate region by exon amplification' Proceedings of the 8th international congress of human genetics, Washington, D.C., US; october 6-11, 1991, abstract no. 2383 ---	1-16